doi: 10.1093/nar/gku968

P2CS: updates of the prokaryotic two-component systems database

Philippe Ortet[1,2,3], David E. Whitworth[4], Catherine Santaella[1,2,3], Wafa Achouak[1,2,3] and Mohamed Barakat[1,2,3,*]

[1]CEA, IBEB, Lab Ecol Microb Rhizosphere & Environ Extrem, Saint-Paul-lez-Durance F-13108, France, [2]CNRS, UMR 7265 Biol Veget & Microbiol Environ, Saint-Paul-lez-Durance F-13108, France, [3]Aix Marseille Université, BVME UMR7265, Marseille F-13284, France and [4]Institute of Biological, Environmental and Rural Sciences, Aberystwyth University, Ceredigion, SY23 3DD, UK

Received September 12, 2014; Revised October 1, 2014; Accepted October 02, 2014

ABSTRACT

The P2CS database (http://www.p2cs.org/) is a comprehensive resource for the analysis of Prokaryotic Two-Component Systems (TCSs). TCSs are comprised of a receptor histidine kinase (HK) and a partner response regulator (RR) and control important prokaryotic behaviors. The latest incarnation of P2CS includes 164 651 TCS proteins, from 2758 sequenced prokaryotic genomes.

Several important new features have been added to P2CS since it was last described. Users can search P2CS via BLAST, adding hits to their cart, and homologous proteins can be aligned using MUSCLE and viewed using Jalview within P2CS. P2CS also provides phylogenetic trees based on the conserved signaling domains of the RRs and HKs from entire genomes. HK and RR trees are annotated with gene organization and domain architecture, providing insights into the evolutionary origin of the contemporary gene set.

The majority of TCSs are encoded by adjacent HK and RR genes, however, 'orphan' unpaired TCS genes are also abundant and identifying their partner proteins is challenging. P2CS now provides paired HK and RR trees with proteins from the same genetic locus indicated. This allows the appraisal of evolutionary relationships across entire TCSs and in some cases the identification of candidate partners for orphan TCS proteins.

INTRODUCTION

Two-component signal transduction systems (TCSs) are found across all three domains of life and allow adaptive responses to changes in environmental conditions (1). They are mainly found in prokaryotes, where they control diverse and important behaviors (2). TCSs are composed of two proteins, a histidine kinase (HK) and a partner response regulator (RR), which interact physically with each other, and which are often encoded next to each other in genomes (3). The P2CS database provides easy identification, characterization and analysis of the TCS proteins encoded in prokaryotic genomes. The latest version of P2CS contains 164 651 TCS proteins from 2758 prokaryotic genomes and can be accessed at http://www.p2cs.org/.

TCS are a diverse group of signaling pathways, which makes their constituent proteins difficult to characterize by sequence analysis (4,5). HKs can be easily identified by possession of a transmitter domain/module and RRs are defined by possession of receiver domains. During signal transduction, phosphoryl groups are transferred directly between transmitter and receiver domains. However, TCS proteins are usually multi-domain proteins and have diverse domain architectures, including various sensory and effector domains (4,6,7). Although partner HKs and RRs are typically encoded by adjacent genes, isolated HK and RR genes (orphans) are abundant, as are complex gene loci encoding multiple TCS proteins (3). Hybrid HK genes are also numerous and can encapsulate entire TCS signaling pathways and relays in single proteins (8). The structures of TCS signaling pathways are defined by the interactions between their transmitter and receiver domains, but these are often not apparent from genome sequences except in the simplest cases. Consequently, there is a need for tools, which shed light on the gene, protein and domain partnerships of TCSs.

P2CS is a resource for the TCS research community, providing consistent identification and annotation of TCS genes and their products. P2CS users are able to find TCS proteins of interest through a variety of routes, including domain architecture searches, and browsing through genomes. A full description of each TCS protein is provided, with links to analysis servers for downstream investigations. A 'gene cart' facility allows users to keep track of

[*]To whom correspondence should be addressed. Tel: +33 442 25 63 71; Fax: +33 442 25 48 48; Email: mohamed.barakat@cea.fr

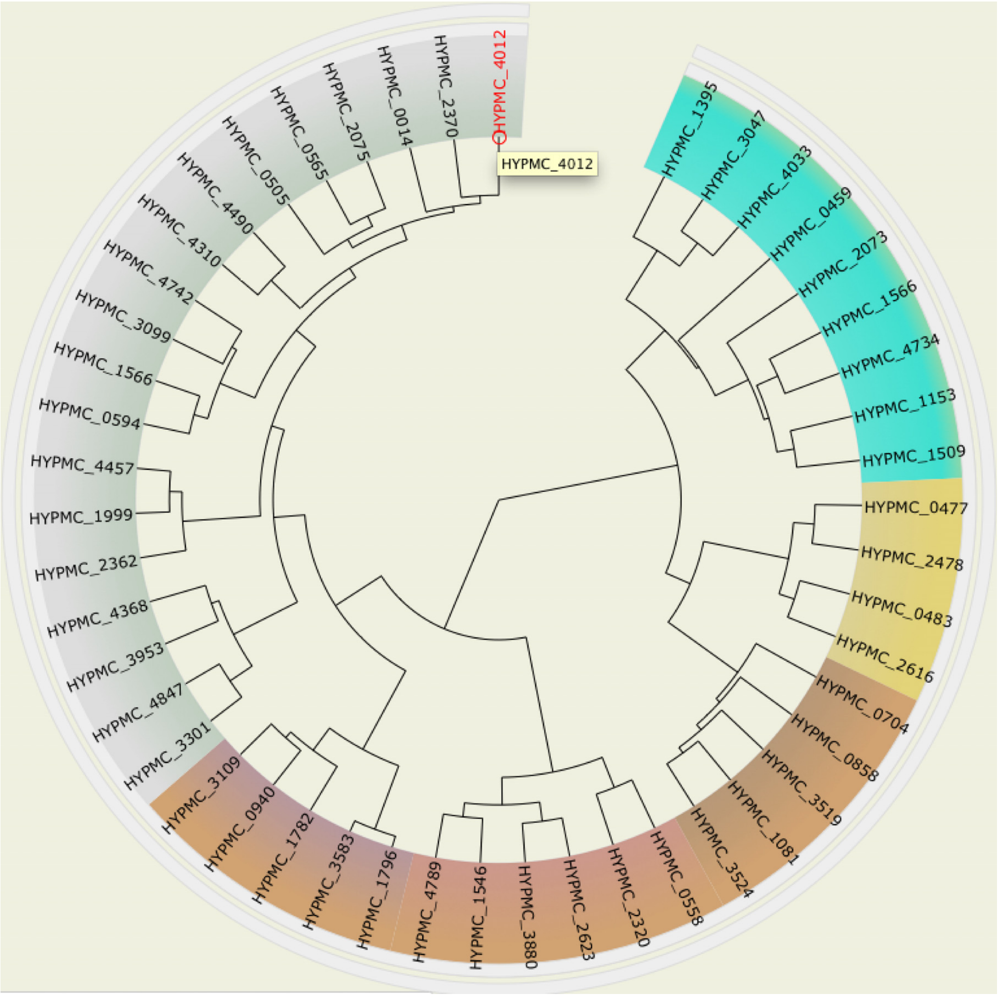

Figure 1. Circular dendrogram from *Hyphomicrobium sp.* MC1. Each locus tag on the dendrogram provides a clickable link to a detailed gene description page.

their proteins of interest and users can download their cart in FASTA, .xls or tab-delimited formats.

The P2CS database was first made public in 2008 (9) and has since been accessed by users nearly 20 000 times. New features and functionalities are regularly added to P2CS, some of which were described in 2011 (10), and it has spawned related websites, including the P2TF database of transcription factors (11) and the P2RP web server, which identifies and characterizes transcription factors and TCS proteins in user sequence queries (12). This article describes novel features of P2CS added since its last description in 2011.

THE P2CS DATABASE IN 2014

P2CS has continued to grow with the increase in the number of publicly available completely sequenced prokaryotic genomes. It currently contains 164 651 TCS proteins from 2758 genomes, more than doubling its size over the last three years. In addition to updating with new genomes, new features have been introduced into P2CS. These features allow users to investigate the evolution and sequence relationships between selected TCS proteins, the results of which can potentially provide information regarding partnerships between TCS proteins.

BLAST queries of P2CS

Users can query P2CS entries by browsing genomes and sequence clusters, by searching gene names/locus tags, and by searching for protein domain combinations. A new feature now allows users to identify homologous of proteins of interest within P2CS by using basic local alignment search tool (BLAST) (13). Users have full control over BLAST parameters/output if they want to vary default options, and can add hits to their P2CS gene cart. P2CS proteins can also be used as BLAST queries against Uniprot or Genbank. The BLAST option can be found under 'Search' in the top bar of each P2CS page.

Multiple sequence alignments

Homologous proteins (either generated by BLAST queries, from domain architecture searches or by selecting sets of proteins of the same family on a P2CS genome page) can be

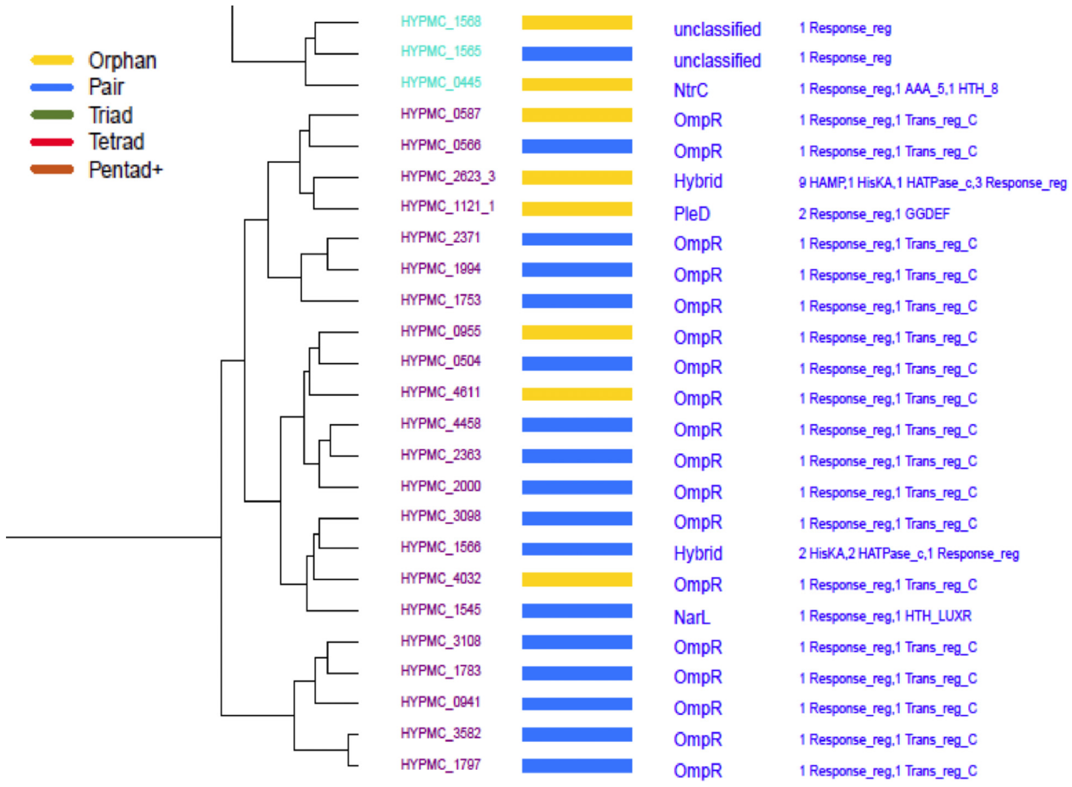
Orphan
Pair
Triad
Tetrad
Pentad+
HYPMC_1568 unclassified 1 Response_reg
HYPMC_1565 unclassified 1 Response_reg
HYPMC_0445 NtrC 1 Response_reg,1 AAA_5,1 HTH_8
HYPMC_0587 OmpR 1 Response_reg,1 Trans_reg_C
HYPMC_0566 OmpR 1 Response_reg,1 Trans_reg_C
HYPMC_2623_3 Hybrid 9 HAMP,1 HisKA,1 HATPase_c,3 Response_reg
HYPMC_1121_1 PleD 2 Response_reg,1 GGDEF
HYPMC_2371 OmpR 1 Response_reg,1 Trans_reg_C
HYPMC_1994 OmpR 1 Response_reg,1 Trans_reg_C
HYPMC_1753 OmpR 1 Response_reg,1 Trans_reg_C
HYPMC_0955 OmpR 1 Response_reg,1 Trans_reg_C
HYPMC_0504 OmpR 1 Response_reg,1 Trans_reg_C
HYPMC_4611 OmpR 1 Response_reg,1 Trans_reg_C
HYPMC_4458 OmpR 1 Response_reg,1 Trans_reg_C
HYPMC_2363 OmpR 1 Response_reg,1 Trans_reg_C
HYPMC_2000 OmpR 1 Response_reg,1 Trans_reg_C
HYPMC_3098 OmpR 1 Response_reg,1 Trans_reg_C
HYPMC_1566 Hybrid 2 HisKA,2 HATPase_c,1 Response_reg
HYPMC_4032 OmpR 1 Response_reg,1 Trans_reg_C
HYPMC_1545 NarL 1 Response_reg,1 HTH_LUXR
HYPMC_3108 OmpR 1 Response_reg,1 Trans_reg_C
HYPMC_1783 OmpR 1 Response_reg,1 Trans_reg_C
HYPMC_0941 OmpR 1 Response_reg,1 Trans_reg_C
HYPMC_3582 OmpR 1 Response_reg,1 Trans_reg_C
HYPMC_1797 OmpR 1 Response_reg,1 Trans_reg_C

Figure 2. A section of the P2CS receiver domain dendrogram output for *Hyphomicrobium sp.* MC1. Bars beside each locus tag show gene organization, as defined previously (10) and colored according to the figure legend. Locus tags are colored according to similarity as defined in the text and the domain architecture and family membership is also indicated alongside the dendrogram.

used to generate multiple sequence alignments using MUltiple Sequence Comparison by Log- Expectation (MUSCLE) (14). Alignments can also be generated for the proteins in the gene cart, which the user can populate with proteins of interest in a diversity of ways as described above. The Jalview applet (15) then allows viewing, editing, coloring and outputting of the alignments within P2CS, for insights into amino acid conservation. The alignments can also be used to generate bespoke phylogenetic trees with the flexible display options of Jalview, shedding light on the evolution of those TCSs.

Phylogenetic analysis

Another novel feature that we have implemented recently is the provision of paired phylogenetic trees, for evolutionary analysis and identification of candidate partners for orphan TCS proteins. For each genome containing more than 6 HKs and 6 RRs, 'Phylogenetic Trees' of HKs and RRs are available in two separate dendrograms, and the corresponding similarity matrices can also be downloaded as 'Heatmaps' or flat texts in various formats. To generate dendrograms, alignments of the phosphotransfer of HKs and receiver domains of RRs and hybrid or unorthodox HKs are constructed using Theorem of the Upper Limit of a score Probability (TULIP) (16), which uses the Smith–Waterman method, with 1000 sequence shuffles, to estimate pairwise Z-values and infers a distance matrix. In R environment - a hierarchical cluster analysis is undertaken using the distance matrix and the hclust function with 'Ward' as a linkage method. An in-house R function then plots the dendrograms to display the hierarchical relationship between domains.

Domains are color-coded by similarity in the dendrograms. Domains of the same color have <60% intra-cluster variance, based on the k-means algorithm. Heatmaps are drawn in the R environment using the heatmap.2 function (with default parameters) and the distance matrix as input. Circular dendrograms generated using the jsPhyloSVG library (17), have clickable locus tags for ease of follow-up navigation (Figure 1), while pdf format trees also provide the gene organization and domain architecture of each protein (Figure 2).

Paired dendrograms

The 'TCS Organization' section of each genome page allows the download of a 'Partnership View', which juxtaposes the HK and RR dendrograms for the organism's HKs and RRs (Figure 3). In the dendrograms, HKs and RRs that are encoded by adjacent genes, are linked together by blue lines and red lines link hybrid or unorthodox HKs to their internal receivers. The TCS gene loci are color-coded by whether they are part of pairs, triads, tetrads etc. as defined by Williams and Whitworth (3). Clades of proteins are color-coded by similarity as mentioned above. As the dendrograms are based on phosphotransfer domain sequence relationships, the trees provide information on evolutionary

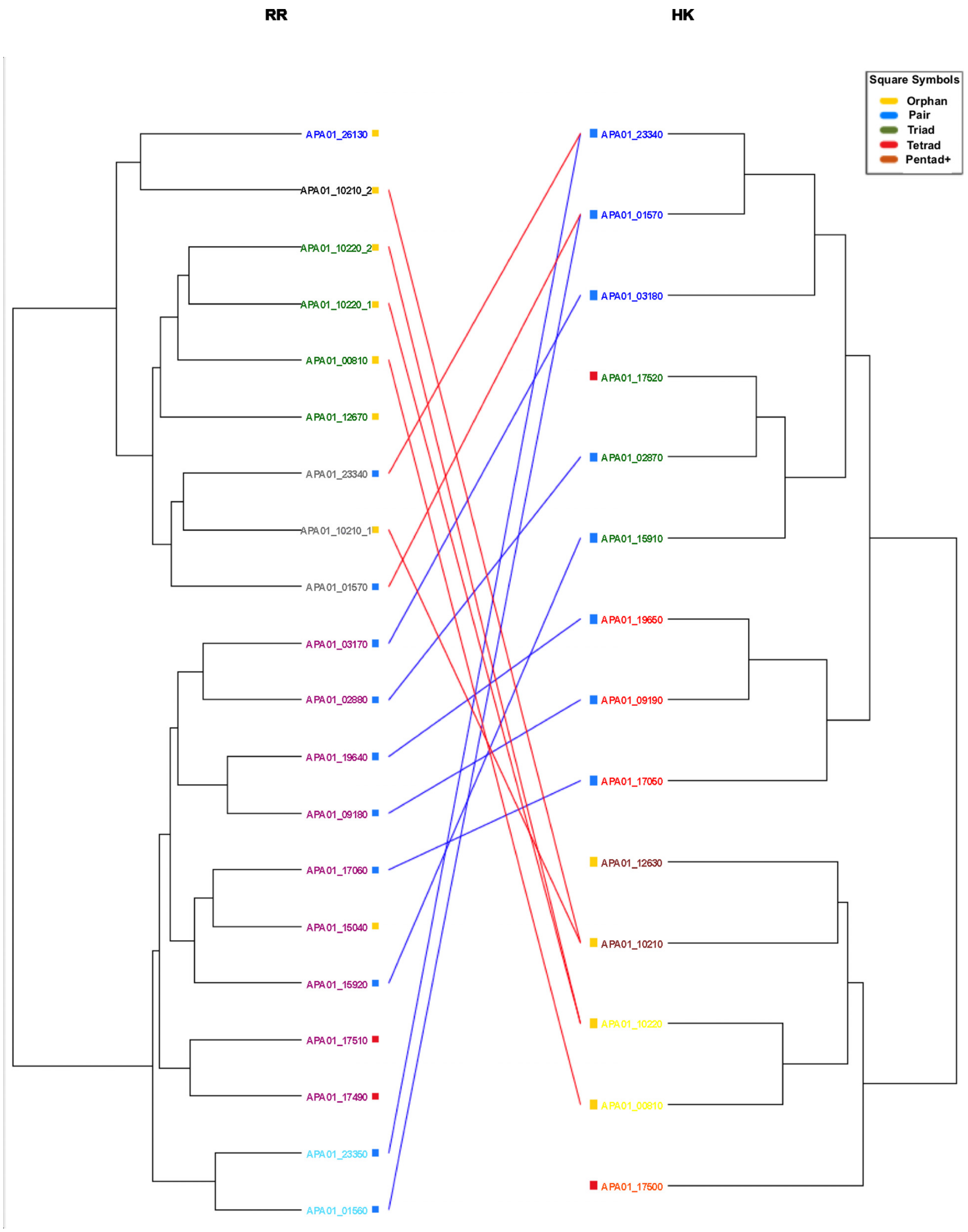

Figure 3. Paired dendrogram from *Acetobacter pasteurianus* IFO 3283–01. The tree of HK sequences (right) is presented opposite that of the receiver domains. Proteins that are encoded by paired genes are joined by purple/blue lines. Red lines connect domains within single hybrid or unorthodox HKs. Red squares next to locus tags indicate that the gene belongs to a locus containing four TCS genes (tetrad organization), blue squares indicate the domain belongs to one of a pair of TCS genes, while a yellow square denotes the gene is orphan.

relationships, but also the interactions between transmitter and receiver domains as suggested for *Bacillus* in an interspecific study (18). This in turn can provide clues into the likely partnerships between proteins and two such examples are described below.

Expanded features

Additional significant improvements to the existing functionalities have been implemented. Each individual gene page now provides an enriched secondary structure highlighting the protein domains, extended clustering data to include clusters sharing 80% identity, and available gene associated literature through UniProt and Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) (19,20) and a link to the PubMed database. Finally, the criteria governing the analysis of genetic organization of TCS genes have been extended to consider in addition to co-oriented genes, neighboring divergent genes localized on opposite DNA strands.

DISCUSSION AND FUTURE DIRECTIONS

The P2CS database is a tool for two-component system analysis. It provides a consistent classification of TCS proteins, alongside diverse tools and intuitive navigation for analysis by users. Recent features implemented in P2CS include the provision of phylogenetic trees of TCS proteins, which can provide important evolutionary insights into TCS biology.

Figure 2 shows a section of the receiver domain tree for the TCS proteins of *Hyphomicrobium sp*. MC1. The purple locus tags define a clade of 22 receiver domains from proteins largely belonging to the OmpR family of RR (possessing Trans_reg_C effector domains (21)). Most of the *ompR* family genes are paired with HK genes as indicated by the colored bars, although four are orphans (HYPMC_0587, HYPMC_0955, HYPMC_4611 and HYPMC_4032). In addition, three of the clustered domains appear to have been recruited to form hybrid kinases (HYPMC_2623 and HYPMC_1566) and a double receiver domain RR (HYPMC_1121). Finally, one receiver domain has a different effector domain, classifying it into a different family (NarL) based on domain architecture (HYPMC_1545). This example illustrates several common features of TCS evolution–gene fusion/fission, conservation of protein architecture with gene organization and domain shuffling (22,23).

As mentioned previously, the paired dendrograms available through P2CS also shed light on evolutionary heritage, and can additionally provide clues about protein HK–RR partnerships. For instance, the genome of *Streptococcus mutans* UA159 encodes a TCS designated VicRK (SMU_1517–SMU_1516) and an orphan RR called CovR (SMU_1924). The orphan CovR and paired VicR fall into the same phylogenetic cluster (Supplementary Figure S1), suggesting that they interact with the same/similar HK. Supporting this inference, Stipp *et al.* (24) provide data that support a model in which VicK, together with both VicR and CovR, coordinates cell division and surface biogenesis. In another example, the paired dendrograms of *Bacillus cereus* ATCC 10987 TCS proteins (Supplementary Figure S2) demonstrate that HKs tend to cluster in patterns mirroring clusters formed by their partner RRs. In this organism there are three HKs which form a clade (BCE_3178, BCE_4968, BCE_1587). A matching clade of three RRs is also found (BCE_3620, BCE_3179 and BCE_4969). In each clade, two proteins are encoded by genes paired with members of the opposite clade. The remaining proteins are orphans. The dendrograms thus suggest that the two orphans (BCE_1587 and BCE_3620) are likely partners for one another. In fact, such a partnership has already been suggested by the study of de Been *et al.* (18).

We will continue adding new genomes to P2CS as they are made public, and will continue developing new functionalities of P2CS to provide more analysis options for users. Firstly we intend creating a new module, which will allow users to create dendrograms for a user-selected sets of proteins. This should allow the rational engineering/selection of proteins for introduction into a genome, either to act as potential partners for an existing TCS protein or to minimize their cross talk with existing TCS proteins. We also intend to develop a support vector machine for the categorization of TCS proteins, as an alternative to the algorithm based on RPS-BLAST, which P2CS currently uses (10). Our other current aspiration is to emulate within P2CS recently-developed computational approaches that predict protein–protein interactions between HKs and RRs, such as those of Burger and van Nimwegen (25) or Cheng *et al.* (26), so that non-bioinformatician users can predict interaction partners within sets of TCS proteins that they specify.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank members of the P2CS user community for their valuable feedback on the database. We are grateful to DSV/IBITEC-S/GIPSI team and particularly Arnaud Martel and Jean-Marc Le Failler for the hosting server installation.

FUNDING

UMR 7265 CEA, CNRS, Aix-Marseille Université. Funding for open access charge: UMR 7265 CEA, CNRS, Aix-Marseille Université.

Conflict of interest statement. None declared.

REFERENCES

1. Wuichet,K., Cantwell,B.J. and Zhulin,I.B. (2010) Evolution and phyletic distribution of two-component signal transduction systems. *Curr. Opin. Microbiol.*, **13**, 219–225.
2. Whitworth,D.E. (2012) Two-component regulatory systems in prokaryotes. In: Filloux,A (ed). *Bacterial Regulatory Networks*. Horizon Press, Caister, pp. 191–222.
3. Williams,R.H. and Whitworth,D.E. (2010) The genetic organisation of prokaryotic two-component system signalling pathways. *BMC Genomics*, **11**, 720.
4. Galperin,M.Y. (2010) Diversity of structure and function of response regulator output domains. *Curr. Opin. Microbiol.*, **13**, 150–159.
5. Whitworth,D.E. (2012) Classification and organization of two-component systems. In: Gross,R and Beier,D (eds). *Two-Component Systems*. Horizon Scientific Press, Norfolk, pp. 1–20.
6. Cheung,J. and Hendrickson,W.A. (2010) Sensor domains of two-component regulatory systems. *Curr. Opin. Microbiol.*, **13**, 116–123.
7. Krell,T., Lacal,J., Busch,A., Silva-Jiménez,H., Guazzaroni,M-E and Ramos,J.L. (2010) Bacterial sensor kinases: diversity in the recognition of environmental signals. *Ann. Rev. Microbiol.*, **64**, 539–559.
8. Raghavan,V. and Groisman,E.A. (2010) Orphan and hybrid two-component system proteins in health and disease. *Curr. Opin. Microbiol.*, **13**, 226–231.
9. Barakat,M., Ortet,P., Jourlin-Castelli,C., Ansaldi,M., Mejean,V. and Whitworth,D.E. (2009) P2CS: a two-component system resource for prokaryotic signal transduction research. *BMC Genomics*, **10**, 315.
10. Barakat,M., Ortet,P. and Whitworth,D.E. (2011) P2CS: a database of prokaryotic two-component systems. *Nucleic Acids Res.*, **39**, D771–D776.

11. Ortet,P., De Luca,G., Whitworth,D.E. and Barakat,M. (2012) P2TF: a comprehensive resource for analysis of prokaryotic transcription factors. *BMC Genomics*, **13**, 628.
12. Barakat,M., Ortet,P. and Whitworth,D.E. (2013) P2RP: a web-based framework for the identification and analysis of regulatory proteins in prokaryotic genomes. *BMC Genomics*, **14**, 269.
13. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1991) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
14. Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.
15. Waterhouse,A.M., Procter,J.B., Martin,D.M., Clamp,M. and Barton,G.J. (2009) Jalview Version 2–a multiple sequence alignment editor and analysis workbench. *Bioinformatics*, **25**, 1189–1191.
16. Bastien,O., Ortet,P., Roy,S. and Marechal,E. (2005) A configuration space of homologous proteins conserving mutual information and allowing a phylogeny inference based on pair-wise Z-score probabilities. *BMC Bioinformatics*, **6**, 49.
17. Smits,S.A. and Ouverney,C.C. (2010) jsPhyloSVG: a javascript library for visualizing interactive and vector-based phylogenetic trees on the web. *PLoS One*, **5**, e12267.
18. de Been,M., Francke,C., Moezelaar,R., Abee,T. and Siezen,R.J. (2006) Comparative analysis of two-component signal transduction systems of Bacillus cereus, Bacillus thuringiensis and Bacillus anthracis. *Microbiology*, **152**, 3035–3048.
19. The Uniprot Consortium (2013) Update on activities at the Universal Protein Resource (UniProt) in 2013. *Nucleic Acids Res.*, **41**, D43–D47.
20. Franceschini,A., Szklarczyk,D., Frankild,S., Kuhn,M., Simonovic,M., Roth,A., Lin,J., Minguez,P., Bork,P., von Mering,C. *et al.* (2013) STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res.*, **41**, D808–D815.
21. Finn,R.D., Bateman,A., Clements,J., Coggill,P., Eberhardt,R.Y., Eddy,S.R., Heger,A., Hetherington,K., Holm,L., Mistry,J. *et al.* (2014) Pfam: the protein families database. *Nucleic Acids Res.*, **42**, D222–D230.
22. Capra,E.J. and Laub,M.T. (2012) Evolution of two-component signal transduction systems. *Ann. Rev. Microbiol.*, **66**, 325–347.
23. Whitworth,D.E. and Cock,P.J. (2009) Evolution of prokaryotic two-component systems: insights from comparative genomics. *Amino Acids*, **37**, 459–466.
24. Stipp,R.N., Boisvert,H., Smith,D.J., Hofling,J.F., Duncan,M.J. and Mattos-Graner,R.O. (2013) CovR and VicRK regulate cell surface biogenesis genes required for biofilm formation in Streptococcus mutans. *PLoS One*, **8**, e58271.
25. Burger,L. and van Nimwegen,E. (2008) Accurate prediction of protein-protein interactions from sequence alignments using a Bayesian method. *Mol. Syst. Biol.*, **4**, 165.
26. Cheng,R.R., Morcos,F., Levine,H. and Onuchic,J.N. (2014) Toward rationally redesigning bacterial two-component signaling systems using coevolutionary information. *Proc. Natl. Acad. Sci. U.S.A.*, **111**, E563–E571.